L1: Introduction

Data Science for Digital Health

Hobson Lane, UC San Diego
Instructor

UC San Diego EXTENSION

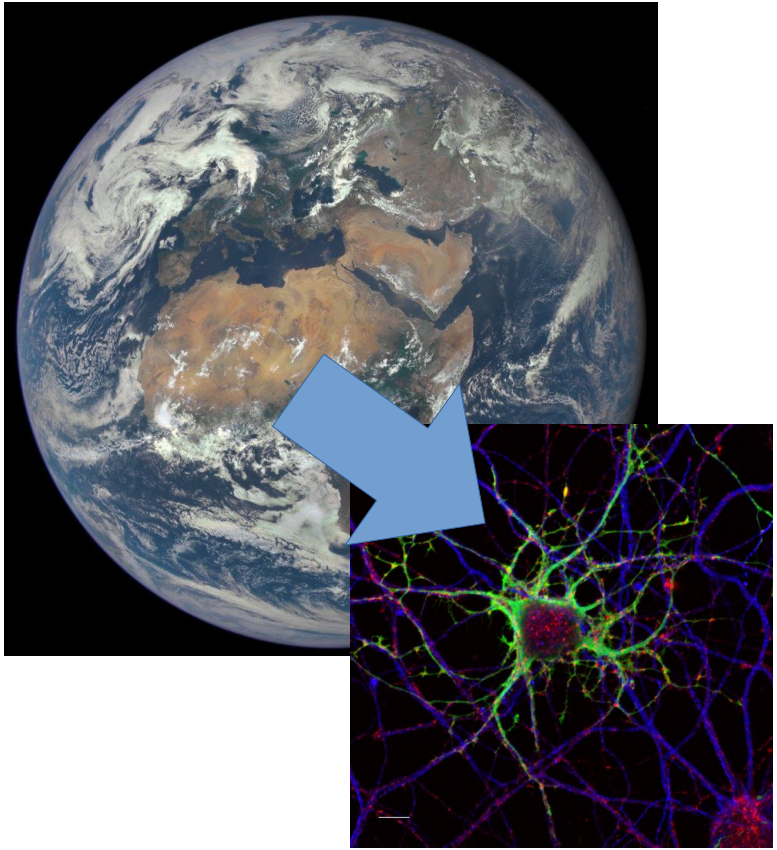
UC San Diego SCHOOL OF MEDICINE

Agenda

Topic	Key Concepts
Approach	Global to personal Big picture to microscopic Practical everyday examples
Data Science for Digital Health	Syllabus
Data Science	Statistics Computer Science Machine Learning AI
Real world application	Kidney disease prediction Breast Cancer
Data science process	Exploring your data Clean your data Evaluate your model

Learning Progression

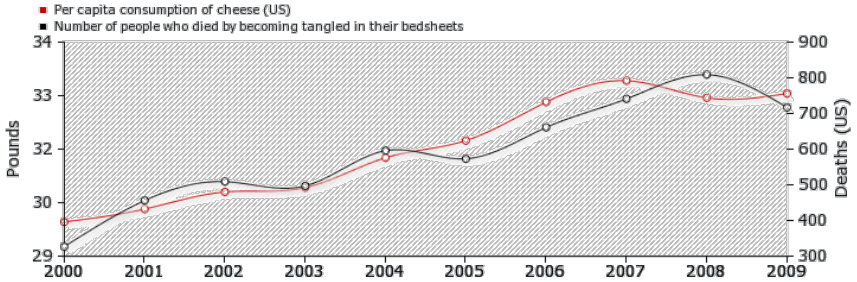
- Big picture → microscopic
- Population → patient
- Theory → example
- Epidemiology → biology



Pedagogy

- History and philosophy first
- Quintessential examples
- Ask questions like Socrates
- Take notes and share
- Learn by teaching
- Critical thinking

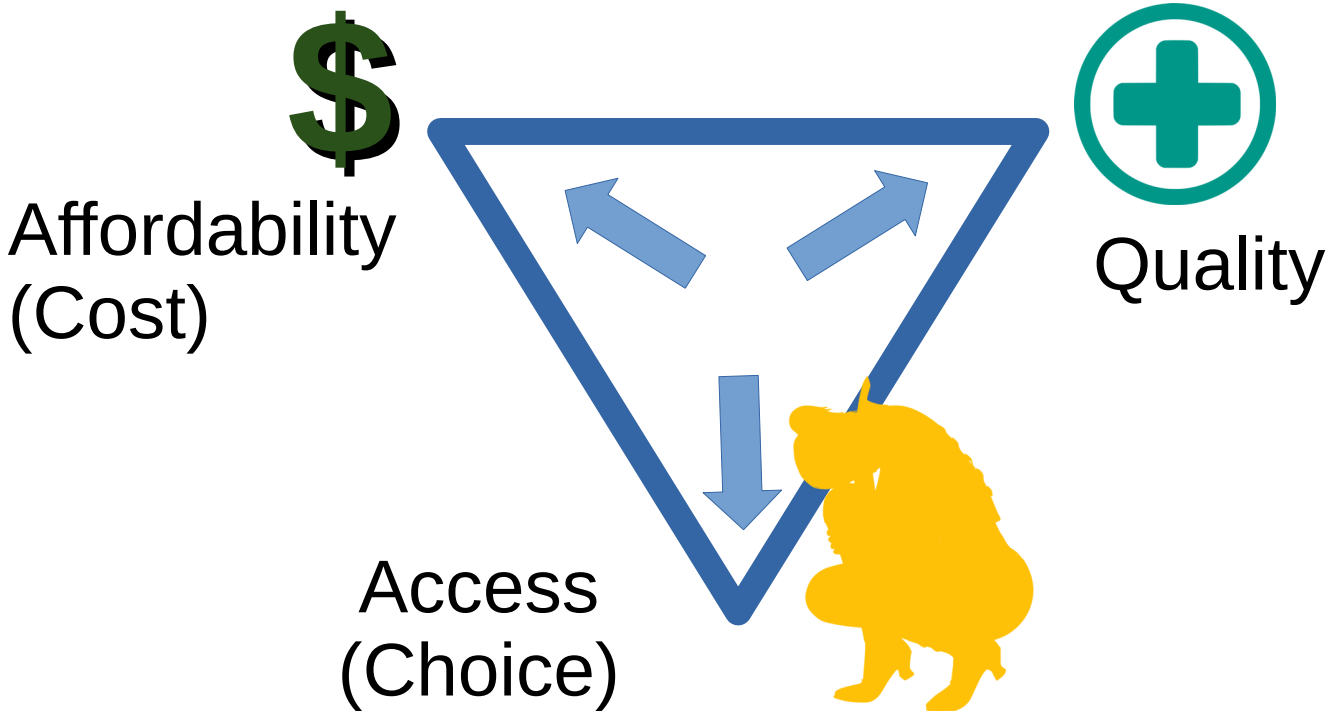
Spurious correlation example



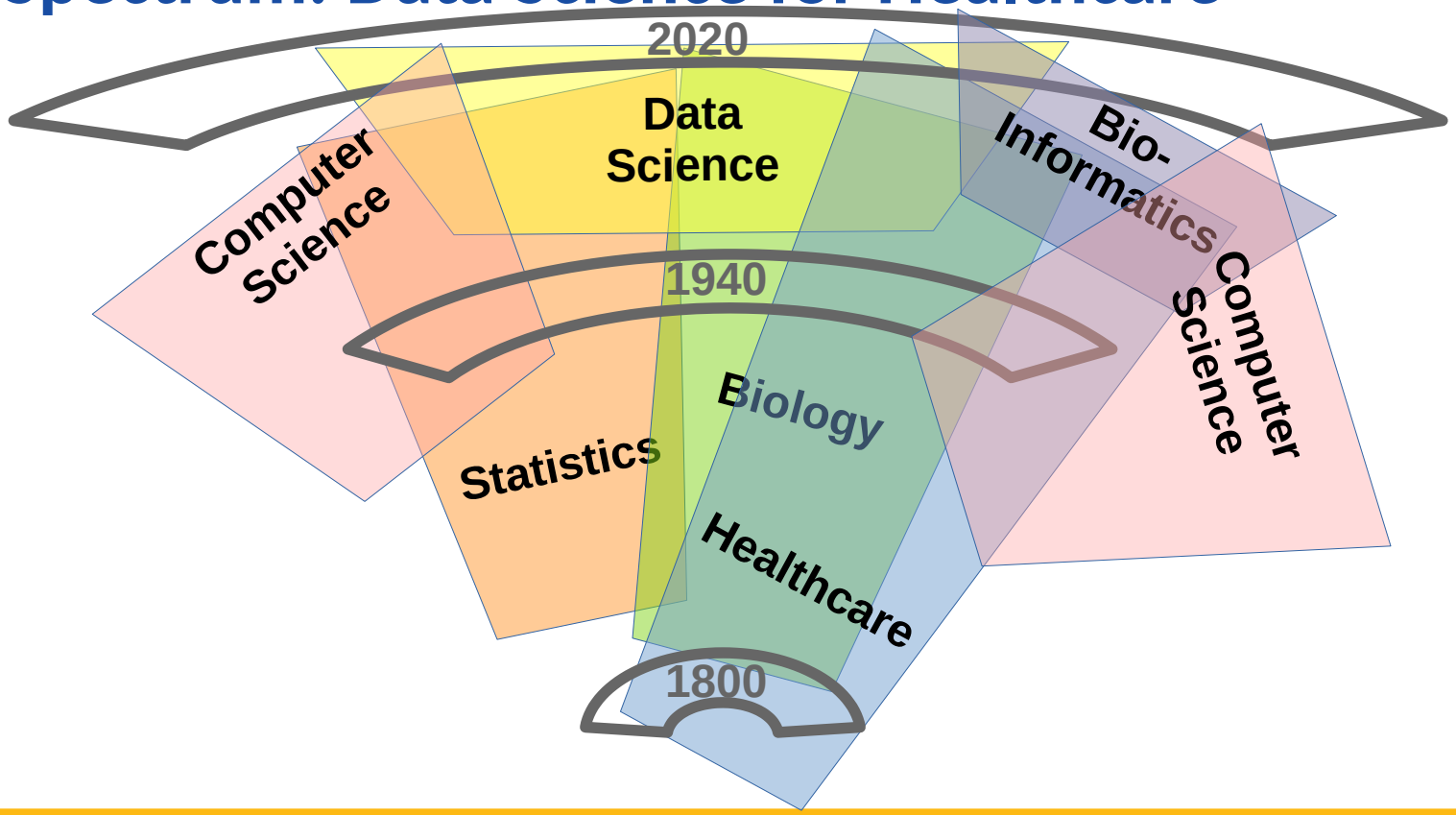
Syllabus with exercises

Wk.	Title	Topics	Exercise
1	Data Science in Healthcare	Application, terminology, pedagogy	anonymize dataset PII
2	Spreadsheet Data Science	ETL, exploration & visualization	height, weight, BMI, gender
3	Statistics and Privacy	PII, causality, MLE	causal diagram “games”
4	Clinical Data Science, Machine Learning	Predictive models, Bayes’ rule	predict diabetes risk
5	Deep Learning & Computer Vision	neural nets, radiology, cv	train diagnostic neural net
6	Natural Language Processing	virtual assistance, summarization, text mining	**Mid-term Quiz!**
7	Bioinformatics and genomics	Alignment, sequences, gene, chromosome, phenotype	sequence modeling
8	Assistive Technology & Ethics	Tesla, Aira, OSHA, Gap Anal.	OCR vs. barcodes for pharma
9	Healthcare Systems Modeling	time series, business models, objective functions	Reduce hospital readmission
10	Public Health & Epidemiology	GIS, spatio-temporal modeling	analyze opioid epidemic
Proj	Train your own healthcare model	find data, ETL, fit & evaluate model	**Fit/train a DS model!**

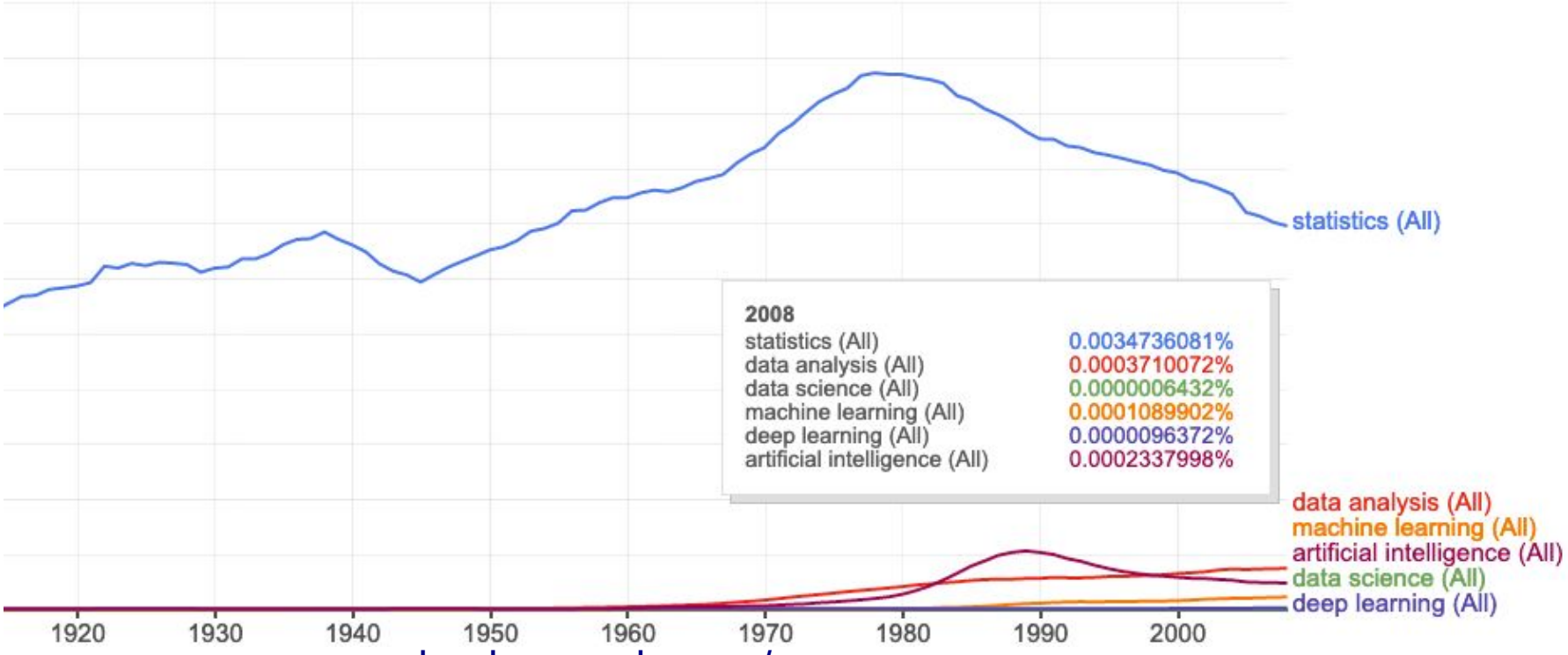
The Iron Triangle (1994)



Spectrum: Data Science for Healthcare

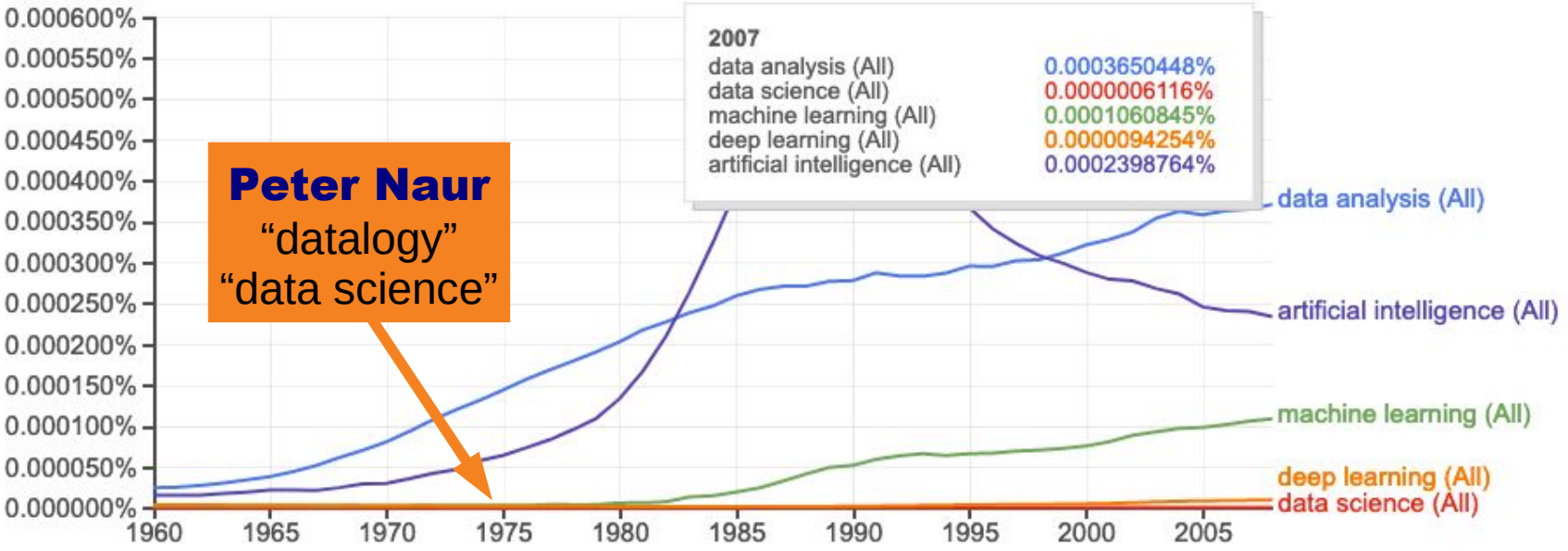


Statistics about the word “statistics”



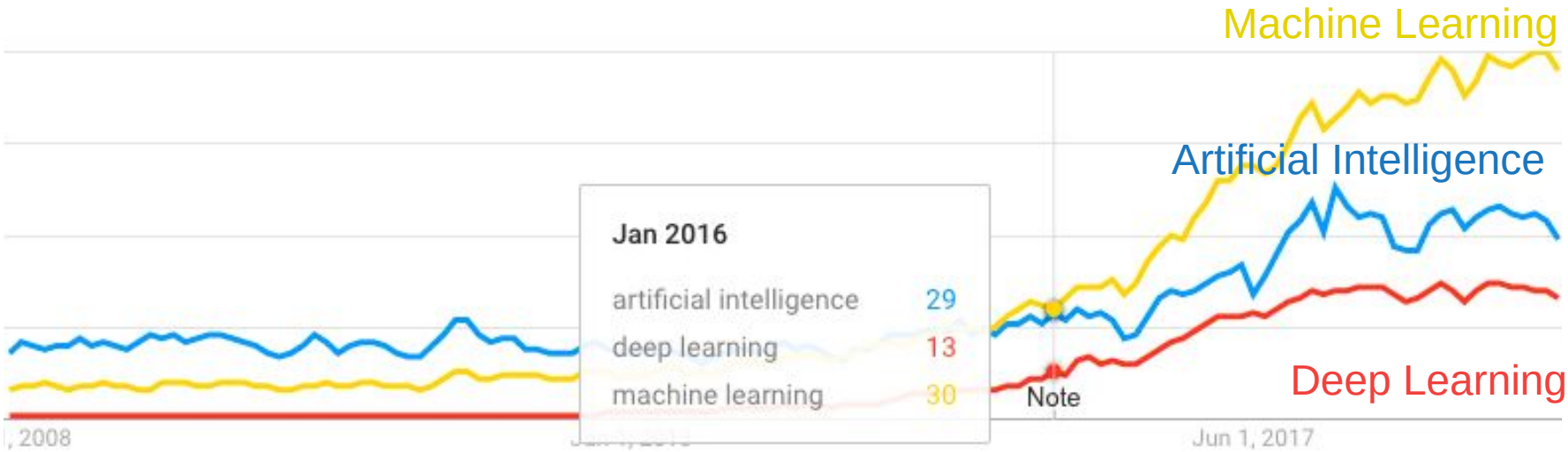
books.google.com/ngrams

Statistics about “data science”



books.google.com/ngrams

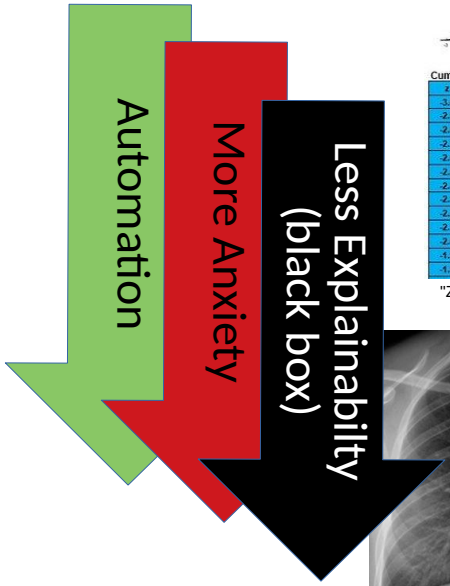
Web search trends

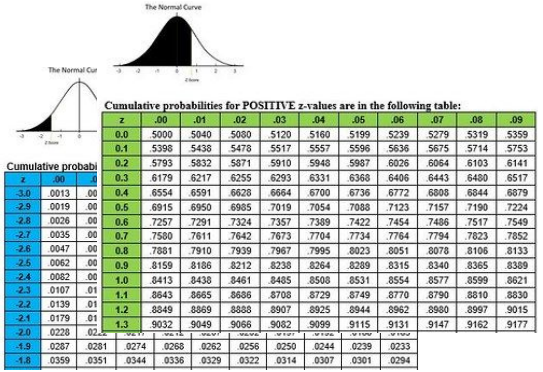


trends.google.com

Automation

- Statistics
- Science
- Data Analytics
- Data Science
- Predictive Analytics
- Machine Learning
- Deep Learning
- Artificial Intelligence (AI)

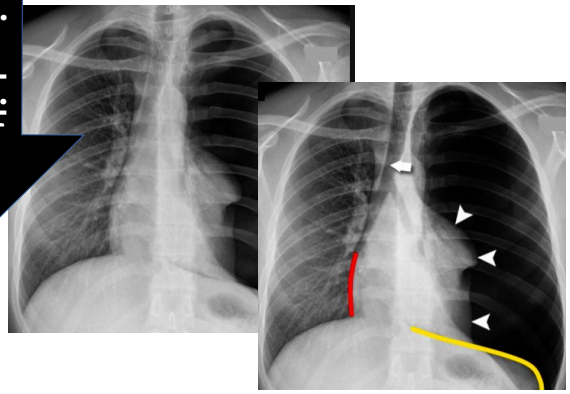




Cumulative probabilities for POSITIVE z-values are in the following table:

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177

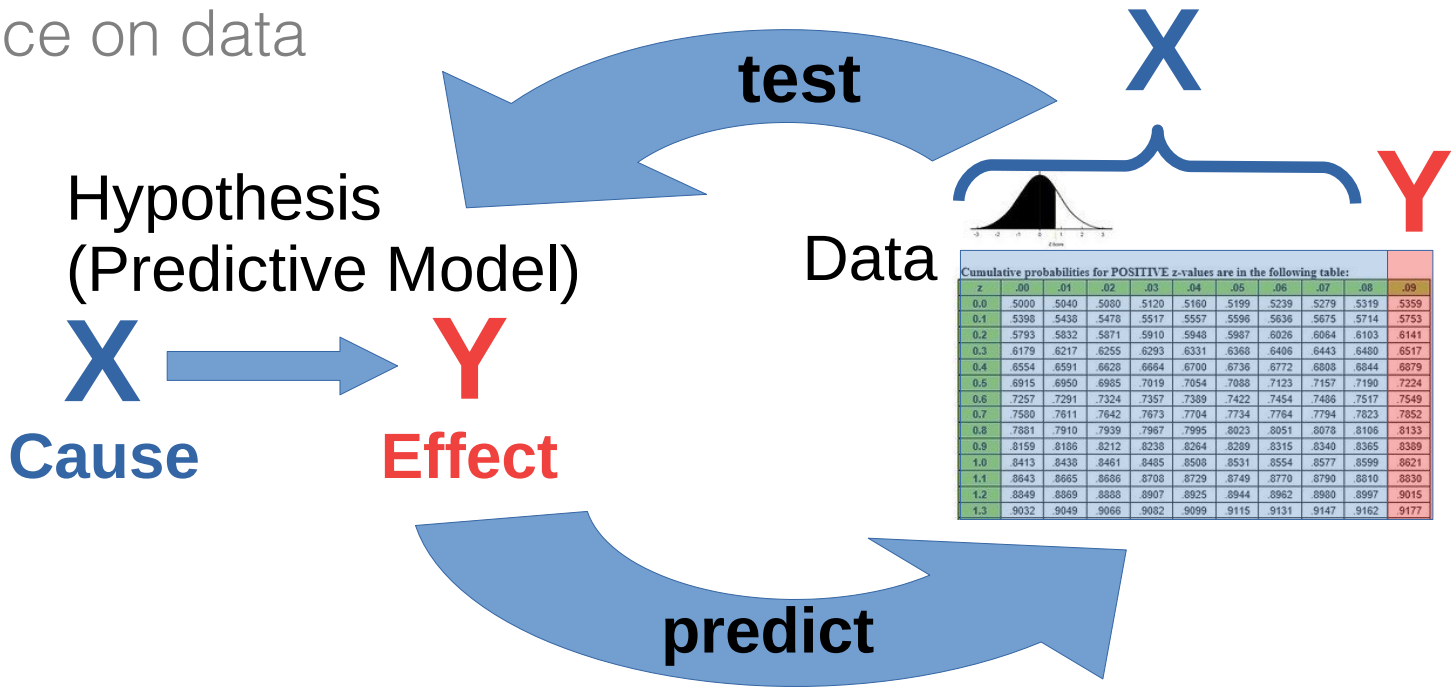
"Z Tables" by sustainablerural is licensed under CC BY 2.0



radiologymasterclass.co.uk

What is Data Science?

science on data



Cumulative probabilities for POSITIVE z-values are in the following table:

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177

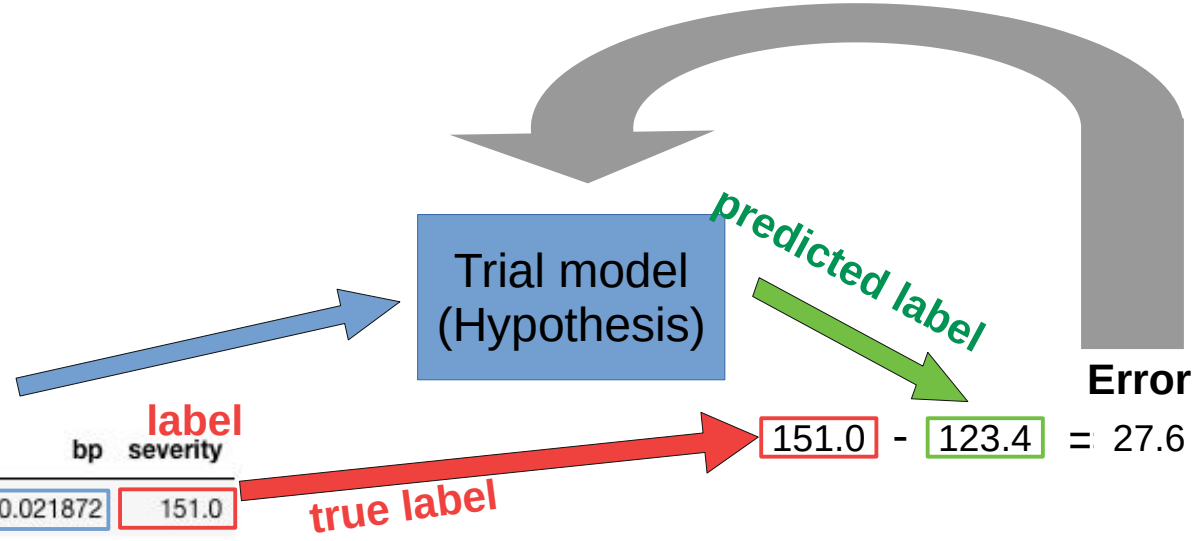
Trial and Error

Diabetes data

features (indicators)

label

	age	sex	bmi	bp	severity
0	0.038076	0.050680	0.061696	0.021872	151.0
1	-0.001882	-0.044642	-0.051474	-0.026328	75.0
2	0.085299	0.050680	0.044451	-0.005671	141.0
3	-0.089063	-0.044642	-0.011595	-0.036656	206.0



Advantages of the Data Science approach

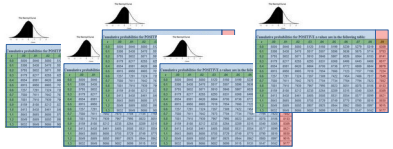
Accidental experiments are often…

Faster

Cheaper

$

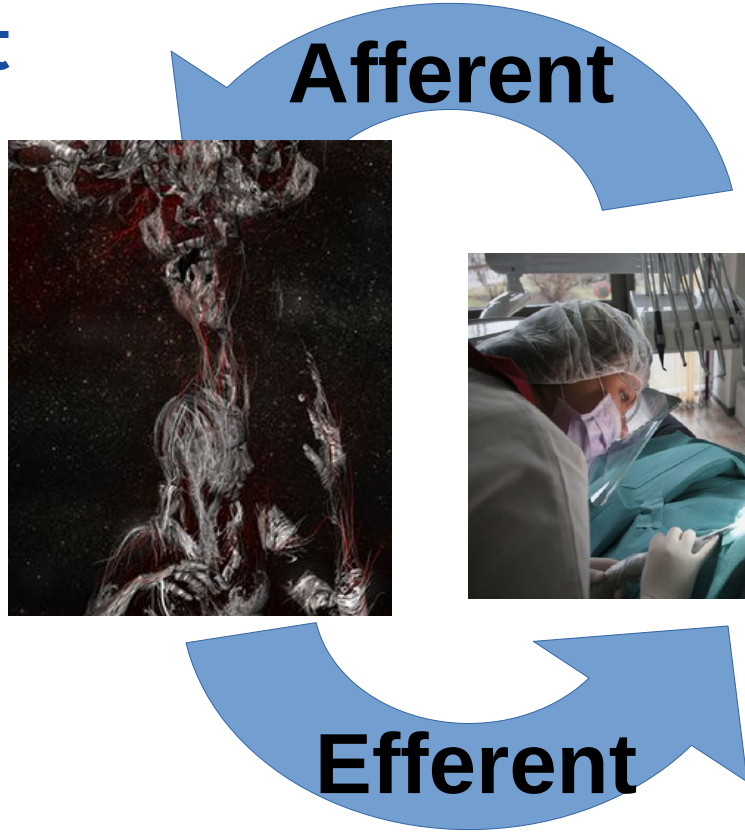
Better



More ethical

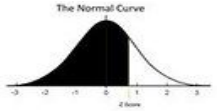
Continuous Improvement

- Learning Health System
  Inst. of Medicine 2006
- Sustainable Learning Sys.
  Charles S. Friedman 2014
- Learning Healthcare Proj.
  Tom Foley, Newcastle Univ



Data science skills

- Extracting and loading data
- Recognizing data types
- Clean and transform data
- Visualizing data
- Modeling data
- Evaluating your model



Cumulative probabilities for POSITIVE z-values are in the following table:

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177

Extracting data

- Text files (CSV, TSV, JSON, TXT)
- Compressed files (ZIP, GZ)
- Binary files (XLS, PDF, Images)
- Web pages (links to HTML)
- Databases

Hint: check out “Pandas” and the `pandas.read_csv()` function

Data types

- **Continuous:** numerical values like height, weight, blood pressure, temperature
- **Categorical:** gender, eye color, disease names
- **Natural language:** symptom descriptions, medical procedure descriptions
- **Sequence:** genome, DNA, RNA, protein, chemical pathways
- **Time series:** treatment timelines, hospital records, EKG/EEG recordings)
- **Geographic:** epidemiology, maps of clinic locations
- **Imagery:** X-rays, MRI slices, CAT scan slices, photos of skin abnormalities

Geographic Data

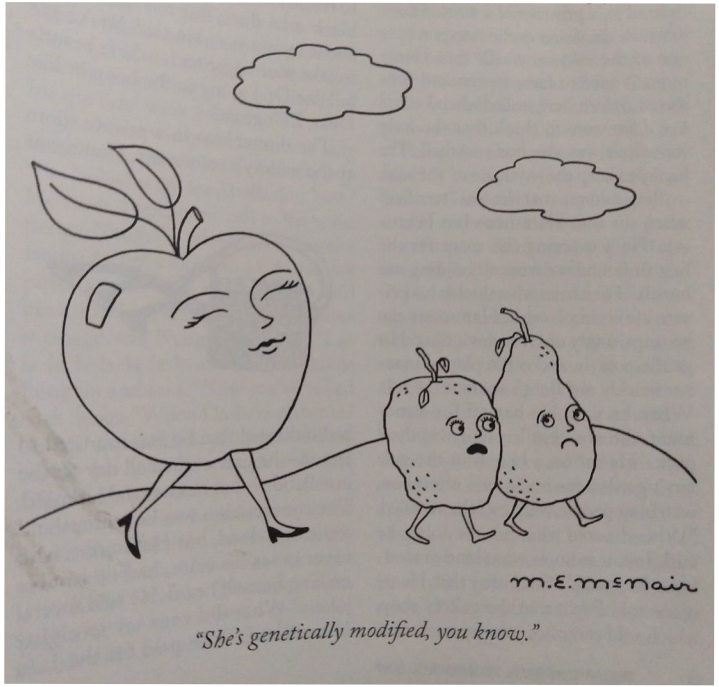
- Pop Health (Population Health)
- Epi (Epidemiology)
  - Annual flu vaccines around the world
  - Hepatitis outbreak in southern US
  - Ebola epidemic in Africa
- Examples
  - Latitude and Longitude
  - State
  - Zip Code

Sequence Data

- Genomics
  - Self-service genetic testing (23andme)
  - Prenatal screening (Counsyl)
  - Pre-exposure allergy prediction
  - Asthma anticipation
  - Resistance to deadly viruses (West Nile, Ebola)

GMO humans

- Prenatal Screening and Genetic Diagnostic Tests
- GMO (selection and breeding)
  - Prenatal screening is effectively GMO of humans
- Reduce disease
  - Unsuccessful CRISPR to cure beta thalassemia in human embryo (2015)
- Resistance to disease
  - Ebola survivors may have resistance encoded in their genome
  - Unsuccessful CRISPR for AIDS resistance in germ line (2016, China)
  - Twins with partial AIDS resistance engineered (2018, China)



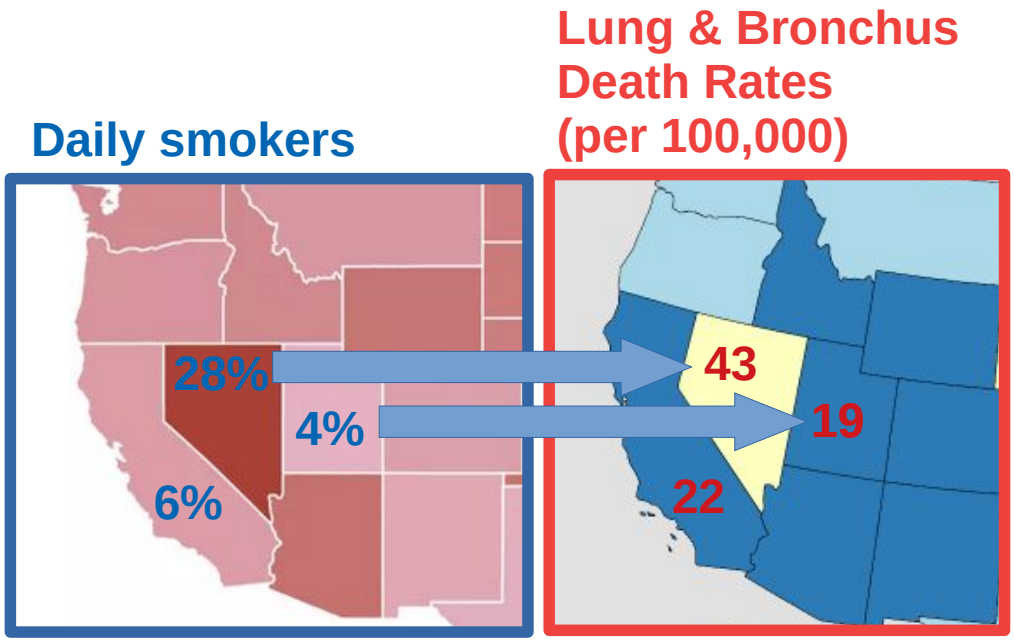
"She's genetically modified, you know."

– M. E. McNair, New Yorker Dec 27, 2019

Gene Editing Ethics

- **Data Science has created a new strain of GMO humans**
  - Prenatal screening
  - GMHs
- **When is it Ethical?**
  - Reduce disease
  - Resistance to disease
  - Ebola survivors may have resistance encoded in their genome
  - Unsuccessful CRISPR for AIDS resistance in germ line (2016, China)
  - Twins with partial AIDS resistance engineered (2018, China)
  - Performance enhancement
  - Radiation resistance to work in nuclear power plants or interplanetary travel?

Smoking → Lung Disease



6	California	22
28	Nevada	43
3	Utah	19

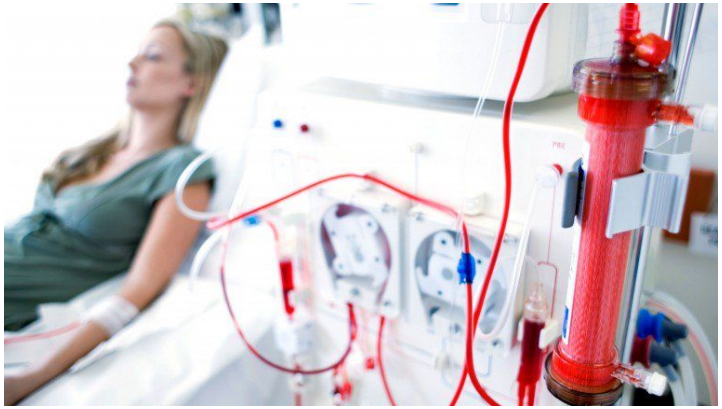
X → **Y**

Smoking **Cancer**

99.98% Correlation

2% P-value

Example Application: Predict Kidney Disease



DeepMind (London)

Clinical records can predict Kidney failure

2 days in advance
55% accuracy for acute problems
90% accuracy for serious issues

Dataset:
100% UK citizens
100% military
90% male

Precision

- Of all the positive results how many were correct?
- Positive predictive value
- True positive rate

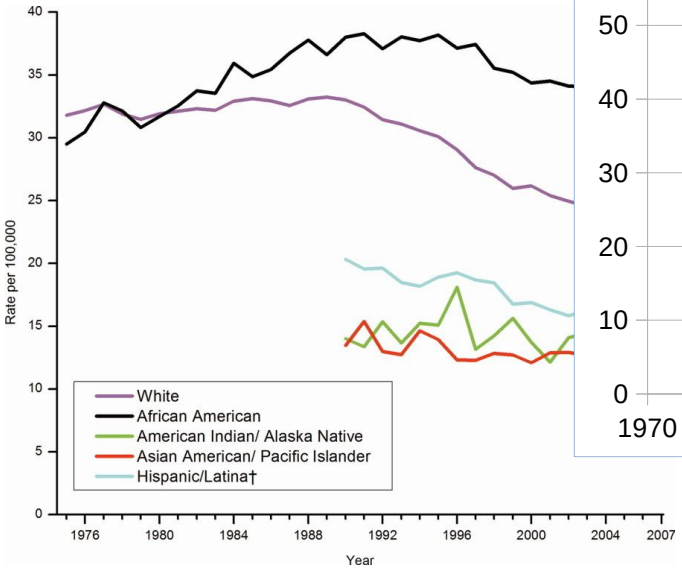
- True_Positive_Count / (True_Positive_Count + False_**Positive**_Count)

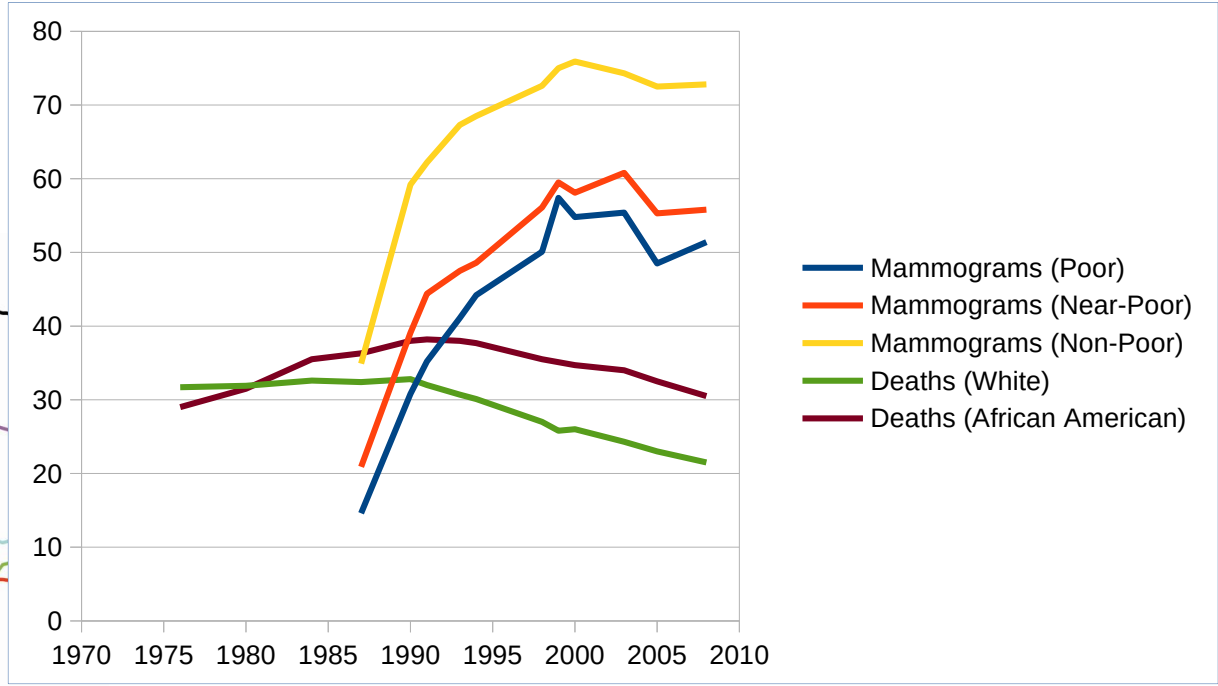
Recall

- Of all the patients with the disease how many were correctly “recalled” (predicted) by the test?
- Sensitivity

- True_Positive_Count / (True_Positive_Count + False_**Negative**_Count)

Correlation enables prediction

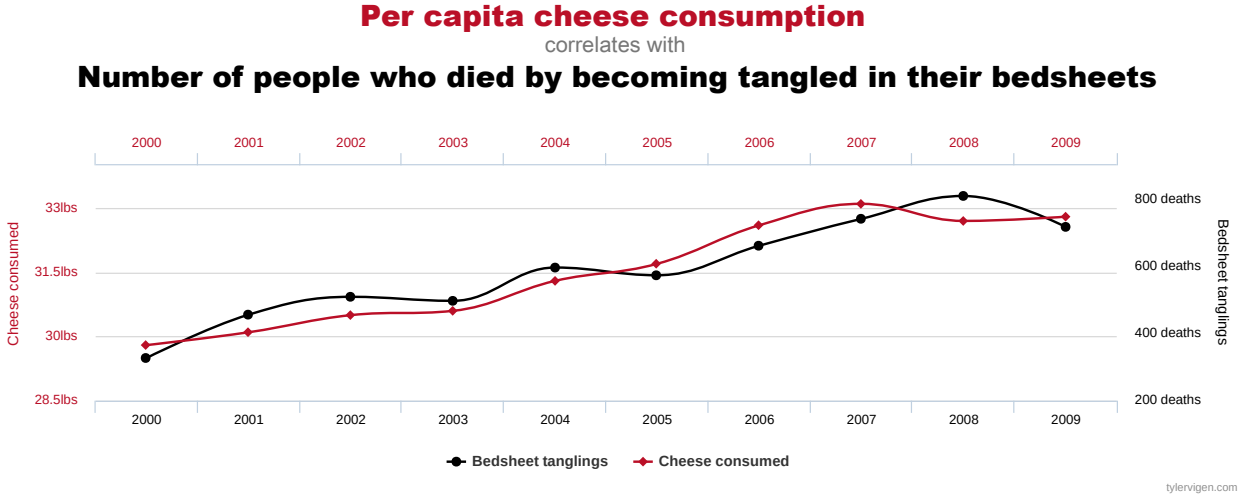




Breast Cancer Rates 2011: bit.ly/ucsdbreast

Correlation is not enough

- Computers are good at finding patterns
- But often those patterns are “spurious correlation”



The “Bradford Hill Criteria”

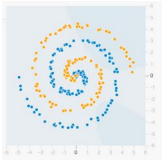
1. Consistency: multiple studies, datasets, subjects
2. Strength: correlation magnitude
3. Dose-response: increased dose = increased effect
4. Specificity: 1 effect rather than many
5. Temporal relationship: cause before effect
6. Coherence: biological plausibility

Assignments

Quiz

1. Give two applications of **Data Science** to Health care
2. How is the **Iron Triangle** of Healthcare?
3. Will **Artificial Intelligence** replace doctors?
4. Why or why not?
5. If a blood test for a particular disease has a False Positive rate of 10% and a False Negative rate of 30%, what's the test's *precision* (positive predictive value) and *recall* (sensitivity)?

Homework: Play with Neural Nets



1. Visit playground.tensorflow.org
2. Select the spiral dataset and add 20% Noise
3. Add and remove different combinations of features: x1, x2, x1^2, x2^2, x1*x2, sin(x1), sin(x2)
4. Play around with different numbers of “HIDDEN LAYERS” and neurons per layer.
5. How many features, hidden layers and total neurons do you need achieve < 15% test set loss?